DOCUMENT RESUME

ED 265 185 TM 860 036

AUTHOR Chissom, Brad; Chukabarah, Prince C. O.
TITLE An Investigation of the Relationship between Item Arrangement and Test Performance.
PUB DATE Nov 85
NOTE 19p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (14th, Biloxi, MS, November 6-8, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; *Comparative Testing; *Difficulty Level; Graduate Students; Higher Education; Item Analysis; *Multiple Choice Tests; *Response Style (Tests); Scores; *Test Format; *Test Items; Test Reliability

ABSTRACT

The comparative effects of various sequences of test items were examined for over 900 graduate students enrolled in an educational research course at The University of Alabama, Tuscaloosa. experiment, which was conducted a total of four times using four separate tests, presented three different arrangements of 50 multiple-choice items: (1) increasing order of item difficulty; (2) descending order of difficulty; and (3) random sequence. Difficulty of the items ranged from .25 to 1.00, and test reliability ranged from .63 to .82 (coefficient Alphas). The scores represented the number of correct items; there was no correction for guessing. No time limits were set, and subjects completed all items. Frequency analyses were conducted, and the descriptive statistics indicated that there were no major differences in mean scores across the experimental replications as well as among the test forms within each replication. Results of the analyses of variance also indicated that the group means were not significantly different. It was concluded that this different arrangement of test items, based on item difficulty, neither relates to power test performance, nor affects examinees' test performance in a positive or a negative manner. However, rearranging test items on different forms may continue to be a useful way to discourage cheating. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED265185

An Investigation of the Relationship Between Item Arrangement and Test Performance

by

Brad Chissom
The University of Alabama

and

Prince C. O. Chukabarah
The University of Alabama

Paper Presented at the Annual Meeting of the
Mid-South Educational Research Association;
Biloxi, Mississippi; Nov.6-8, 1985

NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Chissom, B.

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 860036

INTRODUCTION

Tradition has often required test items to be constructed in an easy-to-hard item-difficulty sequence for various testing purposes. Educators and test constructors adhered to this tradition under the assumption that "students, especially those of average or below average ability, would become quickly discouraged if a test began with difficult test items" (Towle & Merrill, 1975, p. 241), and that students' test-taking anxiety is reduced, thereby facilitating performance (Munz & Smouse, 1968). Even measurement textbooks have suggested easy-to-hard sequencing of testing items because this ascending order of item difficulty pattern purportedly reinforced the examinee positively in "a continuation of purposeful test-taking behaviors" (Kleinke, 1980, p. 225).

Sax and Cromack (1966) observed that there was a dearth of experimental studies dealing with the effects of various types of item arrangements on test-takers. For about two decades now, experiments have attempted to determine the exact relation between test-item arrangement or sequence and the test-takers' characteristics such as achievement examination performance, text anxiety, age, sex, groups/subgroups as well as kinds and nature of a variety of tests such as speeded tests, power tests, mathematics tests, and history tests.

Recently, some researchers have moved from univariate explorations among these variables to multivariate interactional perspectives. But there seems to be no conclusive evidence or consensus for the use of any of the various types of item arrangement, which includes easy-to-hard, hard-to-easy, random, and cyclical forms.

Using the Henmon-Nelson Tests of Mental Ability (Forms A and B), Sax and Cromack (1966) ordered the numerical items according to difficulty levels, arranged them into four different experimental forms (easy-to-hard, hard-to-easy, random, "motivational" cyclical), and administered these forms at random to 276 subjects enrolled in a college introductory course in education. These investigators found that in general the easy-to-hard sequence of item arrangement produced positive significant results if the tests were lengthy or time limits were restricted--that is in speeded tests. No positive significant results were found with respect to power tests or "motivational" tests.

In addition to investigating the effects of anxiety and item difficulty sequence on achievement test scores, Smouse and Munz (1968) examined whether item-difficulty arrangements significantly affected performance scores in different item-difficulty arrangements interacting with test-taking personality factors among 120 male and female subjects enrolled in four sections of an introductory psychology course at a large university. Their results suggested that the use of standard easy-to-hard arrangement of test items is not justified in light of the expected higher performance scores. These authors, therefore, concluded that there exists a more complicated relationship between sequence of test item difficulty, specific personality factors, and achievement test performance.

Marso (1970) conducted two experiments to determine the relationship between test item arrangements and college examinee performance on power tests. This perspective was in contrast to Sax and Cromack (1966) who investigated the effects relative to speeded tests. Based on the results of these experiments, Marso (1970) concluded that "even on relatively lengthy examinations which are

administered as a classroom or power test, one need not be greatly concerned with arranging test items by difficulty, in order of class presentation, or by similarity of content covered at least in terms of influencing required testing time or student performance" (p. 117). Other investigations (Berger et al., 1969; Brenner, 1964; Huck & Bowers, 1972; Kestenbaum & Winer; Smouse & Munz, 1969; Sweeney et al., 1970) have also found no significant differences or other empirical evidence to support the practice of arranging test items in ascending order of difficulty.

The interactive effects of anxiety type and item-difficulty sequencing on mathematics test performance were investigated by Towle and Merrill (1975). The subjects were community college and university student volunteers recruited from mathematics and educational psychology classes. In contrast to the results of Brenner (1964) and Munz and Smouse (1968) in particular, the Towle and Merrill's (1975) results revealed a significant item-difficulty sequence effects, suggesting that items in a speeded test should not be arranged in a hard-to-easy sequence. Although some research studies have reported significant differences (Flaugher et al., 1968; Hambleton & Traub, 1968; Sax & Carr, 1962), it seems that the majority of the research results in the literature reveal no significant differences in the effects of item ordering on examinee's personal characteristics as well as the performance on and the characteristics of the tests themselves.

Under power testing conditions a 96-item examination on a course in Psychiatric Nursing was administered to 104 students in order to investigate the effect of item ordering (easy-to-hard, spiral cyclical, random) and knowledge of ordering and on test scores on students' opinion of the test (Plake, 1980).

The results of this experiment, like many others, revealed no significant differences with respect to all the variables examined. Neither significant order effects, significant knowledge of the orderings nor significant interaction of knowledge of arrangement and order were found. The investigator emphatically concluded that "knowledge, or lack thereof, of item arrangement is not a factor influencing test scores or Student's perceptions about the test for any of the orderings studies" (Plake, 1980, p. 56).

Plake et al (1982) also examined the effects of item arrangement, knowledge of arrangement, test anxiety, and sex on test performance. They used a 48-item multiple-choice mathematics test assembled from items of the ACT College Mathematics Placement Program, and four measures of anxiety. The subjects were 170 students enrolled in three introductory statistics classes; and the instruments were distributed randomly to the students as they arrived for class. Using multivariate analysis of variance the investigators found significant main effects for sex and order, while the sex-by-order interaction was not significant. Females viewed the test as being more difficult. Even when Plake et al. (1983) examined the differential performance of males and females on easy-to-hard item arrangements and the influence of feedback at the item level, they found a significant sex main effect but no significant sex-by-order interaction effect. The female students out-scored the male students when the test items were arranged in an easy-to-hard sequence. However, the researchers cautioned that the lack of a sex-by-order interaction effect could have been due to the use of a nonquantitative examination. The mixed and inconclusive results on the relationship between item order and the sex of the examinee still remains a problem to be resolved.

A slightly different perspective and combination of variables were investigated by Kleinke (1980). He examined the effects of item orders, response location, and sex of the examinee, handedness and pupil's performance on a multiple-choice test. In contrast to many other studies that used college students as subjects, Kleinke's (1980) experiment employed 484 fourth graders; and two item orders were used: easy-to-hard and uniform. Kleinke (1980) found no significant interaction effects, although the main effect of sex was statistically significant with the boys out-performing the girls. Also the left-handed pupils and examinees taking the easy-to-hard tests scored higher. Kleinke's (1980) results like Plake's (1980), Plake et al., (1982) seemed to emphasize the point that item order is an important factor in speeded tests. Plake et al. (1982) citing Mollenkopf (1950) seem to go a step further in saying that "it is well established that, when a test is speeded, the easy-to-hard ordering of the items is best from a psychometric perspective" (p. 55).

Given the various aspects of the relationship between item arrangement and test performance that have been investigated, the present study examined the comparative effects of various test item orderings or sequences on examinees' performances. These effects were examined with graduate students in a large university situation. Therefore, generalization of results of this study to other situations and examination conditions must be made with caution.

METHOD

Subjects

The subjects used in this study were graduate students enrolled in an educational research course at The University of Alabama, Tuscaloosa, during 1982-83 school semesters. The course content included a review of the research process in education from the selection of a problem to the writing of the research report. The instruction consisted of one weekly three-hour class for fifteen weeks presented by permanent members of the teaching faculty. There were some laboratory sessions and small group discussions. Most of the subjects were education majors, and were randomly assigned to their respective testing groups. There were 236 subjects in the first repetition of the experiment, and 286, 223, and 234 in the second, third, and fourth repetitions of the experiment, respectively. The backgrounds of the subjects varied substantially. There were more females than males; and the average age of all subjects was approximately 25, with a range of 22 to 53.

Instruments

The tests for this study consisted of 50 items which were selected and assembled from various test manuals and literature sources. The items were selected to ensure a wide difficulty range, and each test item was structured in a multiple-choice format. Item analysis of the instrument revealed item difficulty indices ranging from .25 to 1.00.

The test items were arranged in three forms, with each form containing the same items but arranged in different item sequence. The first form was an

easy-to-hard sequence, arranged in ascending order of item difficulty. The second form was a hard-to-easy sequence, arranged in descending order of item difficulty. The third, and last form, was a random sequence with the items arranged in a random sequence. The reliabilities of the different forms of the instrument ranged from 0.63 to 0.82 (Coefficient Alphas) as shown in Table 1.

Table 1. Reliabilities (Coefficient Alphas) of the three different forms of the four tests

	Easy-to-Hard Form	Hard-to-Easy Form	Random Form
Test 1	0.76	0.82	0.74
Test 2	0.65	0.72	0.63
Test 3	0.68	0.80	0.78
Test 4	0.79	0.78	0.81

Procedure

For four separate times, the same experiment was conducted with students enrolled in the educational research classes. There were approximately 25 students per class. In administering tests the the subjects were given three forms of the 50-item test at random in the students' classrooms at four separate times during the semester. The four tests covered different topics in the course and did not overlap in content. No time limits were set for the tests and the subjects completed all items.

The collected data, which represented the total number of items correctly answered by each subject, with no correction made for guessing, were subjected to simple frequency analyses and one-way analyses of variance. The single independent factor was item sequencing or ordering, and it had three levels: easy-to-hard, hard-to-easy, and random. The dependent variable was the number of items correctly answered on the test.

RESULTS

The results of this study are presented in three sections as follows:

Descriptive Statistics

The descriptive statistics, computed for all groups, included the sample sizes, means, and standard deviations for each of the four testing times when the experimental replications were conducted. These descriptive statistics were generated for all of the three forms of the test adminstered, and results are shown in Table 2. There were no major differences in the means across the experimental replications as well as among the test forms within each replication. For the four replications of the experiment, the means were 40.22, 40.20, 40.17, and 38.86 respectively. Within each experiment, the means for the different test forms were essentially the same. Consequently, it was no surprise that the results of the analyses of variance (ANOVA) in this experiment indicated that the group means were not significantly different.

Table 2. Descriptive statistics for all groups and experiments.

	Sample	N	Mean	S.D.
Experimental Replication I (Introduction Test)	EASY-TO-HARD	74	41.12	5.22
	HARD-TO-EASY	78	40.36	5.28
	RANDOM	84	39.31	5.96
	TOTAL	236	40.22	5.54
Experimental Replication II (Measurement & Data Collection Test)	EASY-TO-HARD	96	41.05	4.33
	HARD-TO-EASY	105	39.54	5.78
	RANDOM	85	40.04	5.29
	TOTAL	286	40.20	5.21
Experimental Replication III (Library Test)	EASY-TO-HARD	81	40.40	4.17
	HARD-TO-EASY	79	40.11	4.60
	RANDOM	63	39.95	4.04
	TOTAL	223	40.17	4.28
Experimental Replication IV (Historical-Descriptive-Experimental Test	EASY-TO-HARD	83	38.71	5.32
	HARD-TO-EASY	79	38.18	6.13
	RANDOM	72	39.79	4.88
	TOTAL	234	38.86	5.50

Analysis of Variance (ANOVA)

The results of the four one-way analyses of variance revealed that there were no statistically significant differences among the group means ($p < .05$). The respective F-statistics for the four experimental replications were $F(2,233) = 2.136$; $F(2,283) = 2.085$; $F(2,220) = 0.200$; $F(2,231) = 1.750$. These results are shown in Table 4.

These findings are consistent with, and support, the findings of other studies mentioned previously (Kleinke, 1980; Marso, 1970; Munz & Smouse, 1968; Plake et al., 1981; Sax & Cromack, 1966) which, when summarized, indicate that in power tests various item sequences or orderings are not related to examinees' test performances.

Since there was no time limit to complete the tests in the current study, the results may be said to emphasize that in power tests the order or sequence of the test items seem to be irrelevant to the examinees' test performances.

Table 4. ANOVA summary tables for all experiments.

	SOURCE	SS	DF	MS	F
Experimental Replication I (Introduction Test)	FORMS	129.89	2	64.94	2.13
	ERROR	7085.81	233	30.41	
Experimental Replication II (Measurement & Data Collection Test)	FORMS	112.10	2	56.06	2.09
	ERROR	7607.59	283	26.88	
Experimental Replication III (Library Test)	FORMS	7.37	2	3.68	.20
	ERROR	4056.19	220	18.44	
Experimental Replication IV (Historical-Descriptive-Experimental Test)	FORMS	105.18	2	52.59	1.75
	ERROR	6942.45	231	30.05	

Discussion and Conclusion

Under the conditions of this study, no statistical significance was found ($p < 0.05$), indicating that there is no relationship or differential effects between item arrangements and examinee's test performance. These results indicate that test item sequencing in easy-to-hard, hard-to-easy, or random forms, on the bases of item difficulty levels, neither relates to test performance nor affects examinee's test performance in a positive or negative manner. Even though these results agree with those of other studies in the literature, they must be interpreted and applied to situations similar to the ones of this study.

Practically, all the studies, including the present one, that have found no relationships between item sequence and test performance should be of help to test constructors in general and classroom teachers in particular. It may be reasonable to assume that rearranging test items in different orders produces equivalent tests, thus reducing the chances of cheating. For example, in a crowded group or classroom examination condition, there is often the possibility that answers on adjacent student's answer sheets could be copied by his/her neighbor. This possibility is greatly minimized if the items on adjacent copies of the test have been rearranged in various sequences. This would result in a reduction in the opportunity for cheating, and no difference in test performance resulting from different item sequences.

References

Berger, V. F.; Munz, D. C.; Smouse, A. D; & Angelino, H. (1969). The effects of item difficulty sequencing on anxiety reaction type on aptitude test performance. The Journal of sychology, 71, 253-258.

Brenner, M. N. (1964). Test difficulty, reliability, and discrimination as functions of item difficulty order. Journal of Applied Psychology, 48, 98-100.

Flaugher, R. L.; Melton, R. S.; & Myers, C. T. (1968). Item rearrangement under typical test conditions. Educational and Psychological Measurement, 28, 813-824.

Hambleton, R. K.; & Traub, R. E. (1974). The effects of item order on test performance and stress. Journal of Experimental Education, 43, 40-46.

Huck, S. W.; & Bower, N. D. (1982). Item difficulty level and sequence effects in multiple-choice and achievement test. Journal of Educational Measurement, 9, 105-111.

Kestenbaum, J. M.; & Weiner, B. (1970). Achievement performance related to achievement motivation and test anxiety. Journal of Consulting and Clinical Psychology, 34, 343-344.

Kleinke, D. J. (1980). Item order, response location and examinee sex and handedness and performance on a multiple-choice test. Jouranl of Educational Research, 73, 225-229.

Marso, R. N. (1970). Test item arrangement, testing time, and performance. Journal of Educational Measurement, 7, 113-118.

Mollenkopf, W. G. (1950). An experimental study of the effects of item-analysis data of changing item placement and test time limit. Psychometrika, 15, 291-315.

Monk, J. J. (1970). Effects of item order on test scores. Journal of Educational Research, 63, 463-465.

Munz, D. C.; & Smouse, A. D. (1968). Interaction effects of item-difficulty sequence and achievement anxiety reaction on academic performance. Journal of Educatoinal Psychology, 59, 370-374.

Plake, B. S. (1980). Item arrangement and knowledge of arrangement on test scores. Journal of Experimental Measurement, 49, 56-58.

Plake, B. S.; Ansorge, C. J.; Parker, C. S.; & Lowry, S. R. (1982). Effects of item arrangement, knowledge of arrangement test anxiety and sex on test scores. Journal of Educational Measurement, 19, 49-57.

Sax, G.; & Carr, A. (1962). An investigation of response sets on altered parallel forms. Educational and Psychological Measurement, 22, 371-376.

Smouse, A. D.; & Munz, D. C. (1968). The effects of anxiety and item difficulty sequence on achievement testing scores. Journal of Psychology, 68, 181-184.

Smouse, A. D.; & Munz, D. C. (1969). Item difficulty sequencing and response style: A follow-up analysis. Educational and Psychological Measurement, 29, 469-472.

Sweeney, D. J.; Smouse, A. D.; Rupipen, O.; & Munz, D. C.. (1970). A test of the inverted-U hypothesis relating achievement anxiety and academic test performance. The Journal of Psychology, 74, 267-273.

Towle, N. J.; & Merrill, P. F. (1975). Effects of anxiety type and item-difficulty sequencing on mathematics test performance. Journal of Educational Measurement, 12, 241-249.